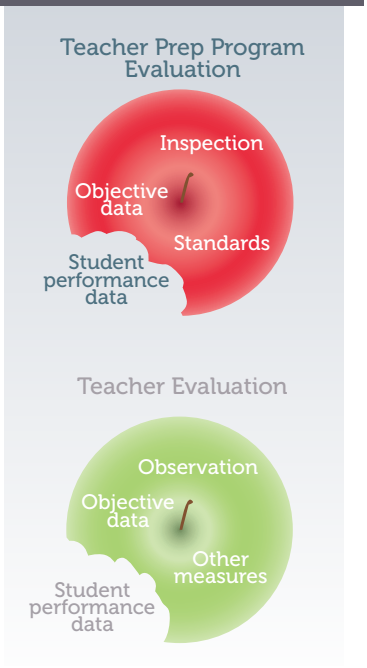
Teacher preparation program student performance data models: Six core design principles

Just as the evaluation of teachers is evolving into a multifaceted assessment, so too is the evaluation of teacher preparation programs. Ideally, evaluation of teacher preparation programs would involve a review of the program against rigorous standards, targeted inspection by objective experts and collection of objective data, of which data on the learning gains of graduates' students are one part. This paper addresses only the use of data on the learning gains of graduates' students to evaluate teacher preparation programs. The use of these data holds great promise because it allows comparison of one program with another in the same state and can help institutions to improve program quality. With this great value, however, comes great challenge.

Because the use of student performance data to evaluate teacher prep is so challenging, states now developing appropriate models might benefit from the experience of early adopters. Here we offer six core principles for strong design based on the models developed in three pioneering states: Louisiana, North Carolina and Tennessee. The principles are outlined below, with a more detailed description of the principles following. While it is possible that current teacher prep data models may be sidelined or at least supplemented in the future as better teacher evaluations become an even richer source of data on student and teacher performance that can be connected back to teacher preparation programs, the same design principles described here will apply to future models as well.

Student Performance Data: One Bite of the Apple



To date, these models have not been formally christened with a name that distinguishes them from the student performance data models that are used to evaluate the performance of individual schools and teachers. To introduce a common nomenclature, we propose that they be called **"teacher preparation student performance data models"** or **"teacher prep data models"** and will use "teacher prep data models" throughout this paper.

1 *Comparisons of institutions across state lines on the basis of results from different state teacher prep data models are not possible at present. The introduction of common student assessments through the PARCC and Smarter Balanced consortia should facilitate some comparisons.*

2 *The purpose of this paper is to address broad design principles, not the statistical fundamentals of the various teacher prep data models. To date, states have selected different statistical models for their teacher prep data models, discussion of which is beyond the scope of this paper. (Louisiana uses a "hierarchical linear model, North Carolina, a "covariate adjustment model" and Tennessee, a "longitudinal mixed effects model.")*

Six Core Principles for the Design and Use of Teacher Prep Data Models

Principle #1:
Data need to be sufficiently specific.
Teacher prep data models should generate findings at the level of specific certification programs within an institution, not just the institution in the aggregate.

Principle #2:
Identifying the "outliers" is what's most important.
The first priority needs to be a model that can accurately identify the value added by relatively larger programs producing graduates who are at the high or low ends of the effectiveness spectrum. This may involve a trade-off between the capacity of teacher prep data models to produce findings for all teacher preparation programs and their capacity to produce actionable findings.

Principle #3:
Use an absolute standard for comparison.
A teacher prep student data model should eventually evaluate how well a program's graduates perform relative to an absolute standard of new teacher performance.

Principle #4:
Try to keep politics out of the technical design of the teacher prep student data model. *The teacher prep student data model is a statistical model, not a political statement, and its design should include the student, classroom and school-level variables that analysis indicates are relevant.*

Principle #5:
Check the impact of the distribution of graduates among the state's K-12 schools.
It is possible that the distribution of graduates among the state's K-12 schools affects the attribution of effects to teacher preparation programs.

Principle #6:
Findings must be clearly communicated.
Teacher prep student data model findings should be explained in reports that are readily comprehensible to policymakers, program administrators and the public at large.

More on the Six Core Design Principles

Principle #1: Data need to be sufficiently specific.

Teacher prep data models should generate findings at the level of specific certification programs within an institution, not just the institution in the aggregate.

Institutions of higher education (IHEs) generally house multiple teacher preparation programs. Our own analysis finds that IHEs house an average of five core elementary, secondary and special education programs at undergraduate and/or graduate levels of training. We have found tremendous variation among these programs. Every aspect of teacher education can vary from program to program, including admission standards, required coursework, the amount of clinical practice and the assigned faculty. Yet, as the table below indicates, for the three state teacher prep data models that now issue public reports, only the model developed in North Carolina is designed to produce findings at the level of a *specific* program rather than of the institution.

The mismatch between how institutions organize preparation and what states' student data models report:

	Louisiana		North Carolina		Tennessee	
	Structure of prep programs	What the teacher prep student data model evaluates	Structure of prep programs	What the teacher prep student data model evaluates	Structure of prep programs	What the teacher prep student data model evaluates
Grade span?	Grades 1-5 Grades 6-12	Grades 4-9	Grades K-6 Grades 6-9 Grades 9-12	Grades 3-5 Grades 6-8 Grades 9-12	Grades K-6 Grades 7-12	Grades 4-8 Grades 9-12
Undergrad, grad or both?	Separate undergrad and grad offered	Only undergrad data	Separate undergrad and grad offered	Only undergrad data	Separate undergrad and grad offered	Combined undergrad and grad data

Only North Carolina does a relatively good job of matching its student data model findings with specific programs. Absent that match, program accountability is impossible.

Given the variation among programs within the same institution, an aggregation of results from numerous programs in one institutional report makes it difficult to ascertain if individual programs are actually producing more or fewer effective teachers. For this reason, a system designed only for institutional accountability is of questionable value.

An explanation for why states are choosing to report a finding at the level of the institution as opposed to the vastly more useful finding at the level of the program is that, at least in the near term, these models would not able to produce reliable findings if they were to further disaggregate. Many institutions simply produce too few graduates in any one program to generate sufficient performance data. As Principle 2 discusses, states should consider that it may be better to produce statistically meaningful results on the few programs that produce sufficient graduates to do so than to produce unreliable results on virtually all programs or institutions.

Principle #2: Identifying the "outliers" is what's most important.

The first priority needs to be a model that can accurately identify the value added by relatively larger programs producing graduates who are at the high or low ends of the effectiveness spectrum. This may involve a trade-off between the capacity of teacher prep data models to produce findings for all teacher preparation programs and their capacity to produce actionable findings.

Expectations about what teacher prep student data model findings can communicate need to be kept realistic. At best, they will be able to distinguish those programs about which one can be fairly confident that graduates are very effective or very ineffective relative to any given standard. Given the many IHEs and other entities involved in teacher preparation within states, it would be difficult to produce a useful value-added measure for every single institution. The reason that this is the case lies in the nature of the statistical analysis as it is applied to data on graduates of a multitude of institutions.

Consider this fact: Over 40 percent of IHEs nationwide produce 50 or fewer teachers in all of their preparation programs combined. The majority of institutions in any given state produce relatively few teachers, making it difficult for some and nearly impossible for others to obtain sufficient data.

Why does data sufficiency matter? The average value added associated with programs' teachers will most certainly vary from program to program in any data set. This variation may indeed reflect persistent average differences in quality across programs. However, there is always the possibility that the variation has been amplified by periods of coincidental "clumps" of either the very effective or the very ineffective teacher candidates that can generally be found in every program. Statistical methods are used to evaluate whether observed differences across sets are due to chance or reflect underlying differences in the preparation of the teachers connected with each set. Differences that are found to be very unlikely to arise due to random chance are called "statistically significant." If actual quality differences across programs are small – as they often are -- a large collection of data is needed to be able to reliably identify differences among data sets that reflect true quality differences among programs and to rule out "accidental results" with some certainty.

To maximize the number of teacher records linked to each preparation program from which data can be obtained, teacher prep data models use a variety of data pooling methods. In Louisiana, for example, four years of successive data are pooled on novice teachers, who are defined as teachers in their first or second year of teaching; North Carolina pools five years of successive data on novice teachers, who are defined as having fewer than five years of experience.

In spite of pooling, production levels can still be so low that no amount of data pooling generates a sufficient number of teacher records for analysis. Because of this, each state has established a "threshold of production" for its teacher prep student data model: Louisiana's model requires a production threshold of 25 teachers per year for inclusion in its teacher prep student data model; North Carolina, 10; and Tennessee, five. In each state, there are a considerable number of small producers who fall below the threshold.

Even for those institutions that have enough teacher graduates to be above the threshold for production and for which their data on graduates will be pooled, the results can still be too statistically imprecise to determine whether graduates really differ from the standard to which they are being compared. For example, 2010-2011 data show that Louisiana State University-Shreveport's graduates produce mean reading scores that are slightly better than the average novice teacher in Louisiana: -1.0 for Shreveport graduates compared to -1.2 for the average novice. But with a difference this small, the results may just reflect random variation. In fact, Louisiana's report on this result indicates that it would take a larger number of graduates than the 38 now produced annually (even using teacher records from four graduating classes and following the performance of each graduate for two years) to have any certainty that the effectiveness of Shreveport graduates is actually different from that of the average novice teacher.

The upshot is that regardless of how data are pooled, no teacher prep student data model can produce the ideal: **reliable results on the effectiveness of a single year's cohort of teachers graduating from every teacher preparation program in a state.**

Teacher prep student data model findings are generally limited to a relatively small subset of programs:

- Large programs with data sets adequate for making a firm conclusion regarding their teachers' performance relative to the state's standard, and
- Programs of any size whose teachers' performance differs so much from the state's standard of comparison that it is possible to draw a statistically reliable conclusion.

Reinforcing a point made at the conclusion of the discussion of Principle 1, decisions that are made about the teacher prep student data model will affect how actionable its findings will be. For example, if more institutions can be included in the teacher prep student data model only by combining undergraduate and graduate program graduates at a given institution, the trade-off in terms of actionable findings may not be worth the additional coverage. Likewise, if more than three years of data on program graduates need to be pooled, the trade-off in terms of actionable findings may be dubious because programs and K-12 school environments may have changed over that time period.

Principle #3: Use an absolute standard for comparison.

A teacher prep student data model should eventually evaluate how well a program's graduates perform relative to an absolute standard of new teacher performance.

Teacher prep data models always produce results about teacher preparation programs relative to one another; results indicate which of the programs or institutions produce graduates that are relatively more effective than others. Moreover, the current standards for comparison are based not on any absolute measure of student progress, but instead on the performance of the average novice teacher in the state, which varies from year to year. The result is that the "best" program in one state may be producing graduates who are less effective than the graduates of the "worst" program in another state. Because there is no way to compare graduates across state lines, it is impossible for any state to know if this is the case.

More interpretive power could be gained from a state's teacher prep student data model if the standard of comparison were instead based on the amount of progress each novice teacher's students should make annually in terms of "normal student learning," perhaps using as a goal college readiness at the end of high school. While it may take some time to set this type of absolute standard appropriately, and there will certainly need to be alignment with the standards in the state's teacher evaluation system, the fact that an absolute standard for novice teachers could be changed as circumstances demand means that states need not let the perfect be the enemy of the good.

Within a few years, the 46 states that have adopted the Common Core State Standards plan to be using one of two sets of standardized tests. Providing that states begin to use more uniform definitions of "novice teacher" (something that now varies among states) and to align their relative or absolute standards of comparison, these assessments will create even more potential to attach interpretative power to student data model results, including interstate comparisons of teacher preparation programs. Ultimately, the capacity to compare preparation programs to one another nationwide, all relative to an absolute standard based on a national conception of annual progress toward college readiness, could produce the most valuable information for teacher preparation improvement and accountability.

Principle #4: Try to keep politics out of the technical design of the teacher prep student data model.

The teacher prep student data model is a statistical model, not a political statement, and its design should include the student, classroom and school-level variables that analysis indicates are relevant.

A variety of variables can be held constant by their inclusion in the teacher prep student data model: student-level variables (e.g., gender, race, level of English proficiency), classroom/teacher-level variables (e.g., percentage of students who are identified as gifted, class mean prior achievement in math) and school-level variables (e.g., percentage of students who receive free or reduced-price meals, school mean prior achievement in reading). The decision of what variables to hold constant while comparing teachers across

preparatory institutions needs to be made based on sound scientific reasoning and experimentation that assesses the degree to which teacher prep student data model results are affected when a particular variable is included. Variables that actually affect student data model results should be included to ensure their proper interpretation. Ideally, no political considerations should enter into decisions about including variables.

How can results be different if a variable is not included? Louisiana, for example, includes the percent of special education students in a classroom as a "classroom variable" in its teacher prep student data model. For each additional one percent of special education students in a classroom, performance is estimated to decrease by about 1.4 percent of a standard deviation. Were this variable to be excluded from the model, the interpretation of the results on the effectiveness of teachers whose classroom differed in their proportions of special education students would be affected: some graduates would look worse than others, but only because they teach a higher proportion of special education students, not because they are truly less effective. Thus, holding constant the share of children who need special education services in each teacher's classroom would help ensure that the report is not placing postsecondary institutions that produce a disproportionate number of teachers whose classrooms have a relatively large proportion of special education children at an unfair disadvantage.

Principle #5: Check the impact of the distribution of graduates among the state's K-12 schools. *It is possible that the distribution of graduates among the state's K-12 schools affects the attribution of effects to teacher preparation programs.*

All current teacher prep data models are considered "value added" because they assess the amount of student academic growth that can be attributed to the teacher in the context of the student, classroom and school variables that can have an impact on student performance. However, some statisticians argue that any variables that are finally included can only account for school characteristics if graduates from the programs being evaluated evenly distribute themselves among different kinds of K-12 schools. If all teachers from one program go to "good schools" and those from another go to "bad schools," these statisticians caution that variables included as theoretical controls for school effects won't actually distinguish whether programs look different because their teachers vary in effectiveness or because their teachers simply manage to find jobs in schools that vary in ways that affect student performance.

To ensure that variables used as school-level controls are effective, teacher prep student data model designers might construct "preparation program networks" by methods described in the technical literature to assess the direct and indirect connectivity of all relevant programs in a window of 2-3 years (sufficient to allow for connectivity, but not so long as to assume that the program and/or the school have not changed). Model designers may recommend that any program that is not directly or indirectly connected in the state's network of preparation programs not be included in the teacher prep student data model.

Principle #6: Findings must be clearly communicated.

Teacher prep student data model findings should be explained in reports that are readily comprehensible to policymakers, program administrators and the public at large.

State's teacher prep student data model reports should not be intelligible only to those well versed in statistics. While technical reports are certainly necessary, of the states currently reporting on its teacher prep student data model, only North Carolina also includes in its set of publications a report that is meaningful to the lay reader. This report translates statistics about program graduates' performance into "days of instruction." In the case of the University of North Carolina (UNC) at Greensboro, for example, the report indicates that graduates add the equivalent of just over two instructional days to their students' performance in elementary mathematics compared to the average North Carolina novice teacher not produced by a UNC system program.

All states' teacher prep student data model reports should, at a minimum, provide the following type of information:

- Most important of all, teacher prep student data model results reported in terms that can be easily understood and compared by the lay reader, such as translating scores into a net gain or loss of instructional days.
- Programs of any size whose teachers' performance differs so much from the state's standard of comparison that it is possible to draw a statistically reliable conclusion.
- The standard relative to which results are reported. For example, in Louisiana, the standard for comparison in mathematics is "the mean adjustment to student outcomes that would be expected" compared to the "mean new teacher effect," which was reported in 2010 to be -3.1. (With the distribution of student outcomes computed to have a standard deviation of 50, this mean new teacher effect represents a decrease in student performance of 6.2 percent of a standard deviation.)
- Clear institutional identification of the groups of graduates about which results are reported: For example, "University of Tennessee, Martin, Undergraduate and Graduate, Grades 4-8" or "Southeastern Louisiana University, Undergraduate, Grades 4-9.
- The IHEs/programs not included in the teacher prep student data model's analysis because they fall below the production size threshold established for the model.
- The IHEs/programs for which results are not statistically significant, making inadvisable a "hard" determination regarding the effectiveness of graduates.

Conclusion: States that have commendably pioneered teacher prep data models have had steep learning curves and have had to make significant adjustments both because of internal kinks and because data systems in K-12 education have been undergoing their own growing pains and adjustments. States now in the process of developing teacher prep data models should learn from the experiences of these states. They should aim to design a student data model that is part of a robust accountability system (including other forms of objective data, standards and inspection) and that will ultimately provide actionable program-specific results based on CCSS-aligned student performance data on the effectiveness of program graduates relative to a standard based on annual progress toward college readiness.

General References

Gansle, K. H., Burns, J. M., & Noell, G., (2010). Value added assessment of teacher preparation in Louisiana: 2005-2006 to 2008-2009 – Overview of performance bands. Louisiana's Teacher Quality Initiative. Retrieved from http://regents.louisiana.gov/assets/docs/TeacherPreparation/200910ValueAddedAssessmentOverviewofPerformanceBandsFINAL82610.pdf

Gansle, K. H., Burns, J. M., & Noell, G., (2011). Value added assessment of teacher preparation programs in Louisiana: 2007-2008 to 2009-2010 – Overview of 2010-2011 Results. Louisiana's Teacher Quality Initiative. Retrieved from http://www.regents.doa.louisiana.gov/assets/docs/TeacherPreparation/2010-11ValueAddedAssessmentOverviewofResultsNov212011.pdf

Henry, G. T., Thompson, C. L., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Marcus, J. V., & Zulli, R. A. (2011). UNC teacher preparation program effectiveness report. Chapel Hill, NC: The Carolina Institute for Public Policy. Retrieved from http://publicpolicy.unc.edu/research/TeacherPrepEffectRpt_Final.pdf

Henry, G. T., Thompson, C. L., Fortner, C .K., Zulli, R. A., & Kershaw, D. C . (2010). The impacts of teacher preparation on student test scores in North Carolina public schools. Chapel Hill, NC: The Carolina Institute for Public Policy. Retrieved from http://publicpolicy.unc.edu/research/Teacher_Prep_Program_Impact_Final_Report_nc.pdf

Mihaly, K., McCaffery, D., Sass, T. R., & Lockwood, J. R. (2012). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. National Center for Analysis of Longitudinal Data in Education Research, Working Paper 63. Retrieved from http://www.caldercenter.org/upload/Mihaly_TeacherPrep.pdf

Noell, G., & Burns, J. M., (2008). Value added teacher preparation assessment Louisiana's Teacher Quality Initiative: Overview of 2007-08 study. Louisiana's Teacher Quality Initiative. Retrieved from http://www.laregentsarchive.com/Academic/TE/2009/2008-09VA(8.27.09).pdf

Noell, G., Burns, J. M., & Gansle, K. H., (2009). Value added assessment of teacher preparation in Louisiana: 2005-2006 to 2007-2008 – Background & new results. Louisiana's Teacher Quality Initiative. Retrieved from http://www.laregentsarchive.com/Academic/TE/2009/2008-09VA(8.27.09).pdf

Tennessee State Board of Education (2008). Report card on the effectiveness of teacher training programs. Retrieved from http://www.tn.gov/sbe/2008Novemberpdfs/II%20A%20Teacher%20Quality%20Report%20Card%20Master.pdf

Tennessee State Board of Education (2009). Report card on the effectiveness of teacher training programs. Retrieved from http://www.tn.gov/sbe/TeacherReportCard/2009/2009%20Report%20Card%20on%20Teacher%20Effectiveness.pdf

Tennessee State Board of Education & the Tennessee Higher Education Commission (2010). Report card on the effectiveness of teacher training programs. Retrieved from http://www.tn.gov/sbe/Teacher%20Report%20Card%202010/2010%20Report%20Card%20on%20the%20Effectiveness%20of%20Teacher%20Training%20Programs.pdf

Tennessee State Board of Education & the Tennessee Higher Education Commission (2011). Report card on the effectiveness of teacher training programs. Retrieved from http://www.tn.gov/thec/Divisions/fttt/account_report/2011reportcard/2011%20Report%20Card%20on%20the%20Effectiveness%20of%20Teacher%20Training%20Programs.pdf

National Council on Teacher Quality

1420 New York Avenue, NW • Washington, DC 20005

Tel: 202-393-0020 Fax: 202-393-0095 Web: www.nctq.org

Subscribe to NCTQ's blog PDQ

Follow NCTQ on Twitter and Facebook

NCTQ is available to work with individual states to improve teacher policies. For more information, please contact:

Sandi Jacobs

Vice President

sjacobs@nctq.org

202-393-0020